Analysis and mitigation of Amazon's face recognition technology in Law enforcement using value sensitive design framework

Akanksha Patil

INFO 542: Foundations of HCI

Dr. Hamid Ekbia

November 29, 2021

Abstract

This research examines algorithmic bias in face recognition tasks led by the Amazon facial recognition algorithm (AFRT) in law enforcement using Value Sensitive Design (VSD) framework. The paper demonstrates the algorithmic bias by discussing the case study of gender shades conducted by authors (Buolamwini & Gebru, 2018) which sheds light on the misidentification of black females by AFRT. The VSD framework is employed to both critically analyse the values in the AFRT algorithm and suggest alternatives that account for certain desired values which will alleviate the misidentifications in law enforcement. In this paper we will look at the key features of the VSD frameworks, understanding algorithmic bias in AFRT algorithm using VSD framework, addressing issues around algorithmic bias in face recognition algorithm, and propose design requirements in existing algorithm based on VSD derived insights and recommendation. In conclusion, the paper

reflects on the benefits and limitations of VSD to analyse the AFRT employed in law enforcement.

Keywords
Amazon facial recognition algorithm (AFRT), Value Sensitive Design (VSD), Facial recognition technology (FRT), American Civil Liberties Union's (ACLU), Value hierarchy table(VHT)

1 Introduction

1.1 Case study: Gender shades

In the study, named “Gender Shades” conducted by researchers at MIT (Buolamwini & Gebru, 2018), various ethical concerns in commercially available facial recognition software were realized. The study involved the collection of a series of photographs and the analysis of the faces was conducted based on gender, facial orientation, and skin complexion. Facial recognition technology (FRT) developed by prominent and infamous organizations like Facebook, Microsoft, and IBM was used to process the photographs. The results of the facial recognition classifiers were seen to perform accurately on male faces as compared to female faces across all three platforms. The error rate in recognizing female faces was in the range of 8.1 percent to 20.6 percent higher than male faces.

On further analysis, it was revealed that the FRT performed worse on black female faces with error rates ranges from 20.8 percent to 30.4 percent (Figure 1). In the years to follow, the notable research findings initiated rapid reactions and stirred the conversation on bias in FRT. Organizations like IBM and Microsoft signalled the revamping of their datasets and user test groups to improve the demographics data collection quality to prevent the bias which was later confirmed in re-audit by gender shades which showed the drop in error rates among black females.

Gender shades further looked into more algorithms, including AFRT, which revealed 31 percent error in gender classification in darker-skinned females (Figure 1) and prompted the existence of racial and gender prejudice. Cases like these have reignited ancient discussions about the relationship between technology and society, highlighting the possibility of automatic discrimination based on traits like age, gender, race, or socioeconomic status. (Winner, 1980)

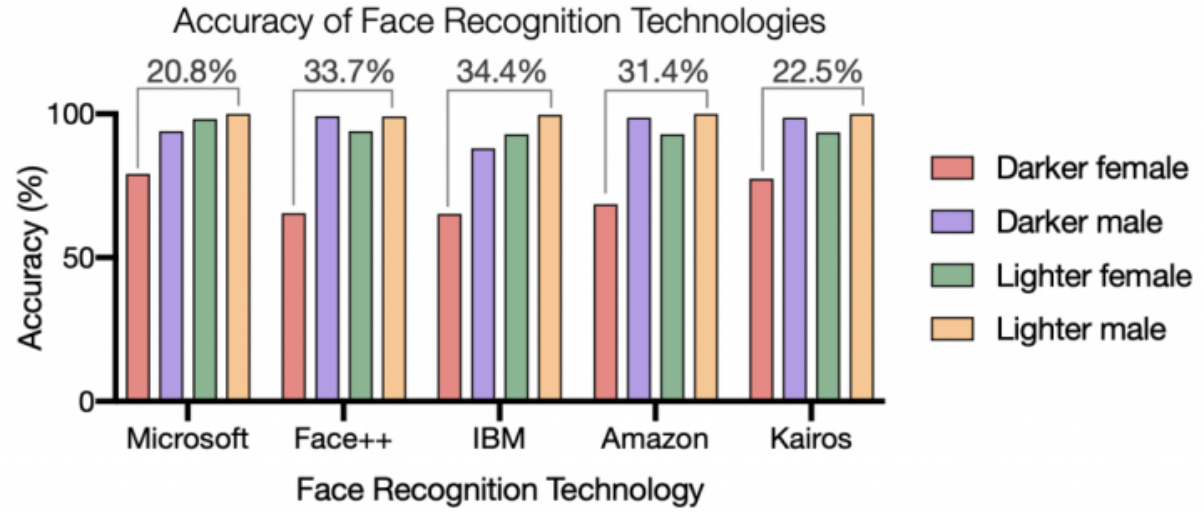

Figure 1: Comparison of five different facial recognition technologies conducted by the Gender Shades research disclosed differences in face recognition technology classification accuracy for different skin tones and sexes. The statistical data has prompted that the algorithms under scrutiny redundantly proved that darker-skinned females had the lowest accuracy and lighter-skinned males had the greatest which resulted into their misidentification.

Source: Najibi A., (2020). Racial Discrimination in Face Recognition Technology. BLOG, SCIENCE POLICY, SPECIAL EDITION: SCIENCE POLICY AND SOCIAL JUSTICE. Retrieved from https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/

1.2 Need for thoughtful design solutions

The potential for automated discrimination casts doubt on neutrality of the algorithm with regards to their ability to structure and shape society rather than simply reflecting on it. However, if technologies are not morally neutral, and the values and disvalues embedded in them have real-world consequences for both individuals and society as a whole, wouldn't this imply that algorithms should be carefully designed and that one should strive to not only detect and analyse problems but to engage with them proactively through thoughtful design decisions?

2 Value sensitive design framework

The aforementioned questions have been debated in the computer science community often and are passed down, even so, they have a long and usually overlooked history in the field.

The Value Sensitive Design (VSD) method, which arose from this conceptual landscape in the mid-1990s (Barocas and Selbst, 2016, p. 673) and has been broadened and refined ever since is the most principled endeavour to design responsibly and sensitively to human values. Interest in the VSD methodology has grown in recent years, owing to a growing awareness that data is not a universal cure and that algorithmic techniques can “affect the fortunes of whole classes of people in consistently unfavourable ways" (Barocas and Selbst, 2016, p. 673) prompting the question: what insights can the approach offer to ongoing debates about bias and fairness in algorithmic decision-making?

The VSD approach, at its core, provides a precise mechanism for actively embedding desirable values into new technology. (Friedman et al.,2006; Flanagan et al.,2008) describe three iterative phases: conceptual, empirical, and technical investigation.

2.1 Conceptual investigation

Conceptual investigation calls for both, recognizing the human values pertaining to the context of investigation and naming the indirect and direct stakeholders. While VSD presents human values relatively broadly as “what is important to people in their lives with a focus on ethics and morality” (Friedman &Hendry, 2019), (Friedman et al. 2006), it is important to follow the specific characteristics of a value defined for specific stakeholders to mitigate the ethical concerns in a sound way. To investigate the conceptual inquiry in the context of FRT in law enforcement, it is important to explore the answers to the questions like,

1. Who are the direct and indirect stakeholders?
2. What values are identified?
3. What values should be chosen?

2.1.1 Who are the direct and indirect stakeholders?

Direct stakeholders are the people who use the technology. In AFRT, direct stakeholders would be the law enforcement officers, owners, developers, and designers of the technology (Keane, 2018). For instance, AFRT is primarily owned by Jeff Bezos which makes him one of the direct stakeholders (Zhou, 2018)
On the other hand, indirect stakeholders are the people who do not use the technology directly but are affected by it. Pertaining to the same, people, in general, are affected due to AFRT, but in particular, people of color are drastically affected due to AFRT.

2.1.2 What values are identified?

The prima facie of the AFRT in law enforcement reveals the bias in the system. But, bias can have a different definition based on the individual and the system. In literal terms, bias means incline. In an undifferentiated context, bias can have a neutral connotation. For instance, if a grocery shopper visits a particular store frequently, the owner of the grocery store can be comparatively biased towards the shopper. But,

bias in moral context could mean discriminating against a group of people. For instance, the grocery store owner could be biased against people who don't belong to his race. In this research paper, algorithmic bias is discussed with respect to case study of AFRT in law enforcement. To identify the varied values emerging under the umbrella of algorithmic bias, categorization is done into 3 abstract categories based on the origin viz., pre-existing, technical, and emerging (Freidman & Nissenbaum,1996).

2.1.2.1 Pre-existing bias can mark the presence in the system explicitly i.e, by individuals perpetrating bias being conscious of their actions while implicit pre-existing bias signals the individuals who are unconsciously subscribing to the biased actions in spite of their good intentions. This can be discretely broken down into an individual and societal standpoint. Individual bias enables the individuals of the system to give inputs that could be biased. For example, in the case of AFRT, some of the employees involved in the processing of user data might be biased towards a particular race, resulting in the formation of the biased dataset. Societal bias enables society at large to deliver its inputs into the system. For example, Amazon's biased recruiting algorithm favoured men over women. (Lauret, 2019)

2.1.2.2 Technical bias arises from technical constraints or considerations. Hardware and tools pose limitations in computer peripherals, software, and hardware that could harbour bias by limiting capacity of machine-learning algorithms thereby creating flaws in the decision making results.

Randomization bias bases the worth of individuals in the system based on the random number allocation. For example, the H1b visa lottery systems pick the individual randomly disregarding the intellectual worth.

Codifying human attributes translates human attributes of judgment, decision making, and instincts into logical data which could be interpreted by computing systems. That is, it follows the quantification of qualities of a human's judgmental construct. For

example, the credit score algorithm favoured white people over black people and the hispanic population because of the prenotion that people from this group would default the interest payments, thus denying them the loans and other benefits.

2.1.2.3 Emergent bias usually occurs after the design is completed as a result of changing societal knowledge, demography, or socio-cultural norms.

New societal knowledge bias emerges because new knowledge emerging in the system cannot be incorporated into the system design. For example, US dollar bills are of the same size, which makes it difficult for visually impaired people to differentiate between them. But, re-design of the bills would impose financial brunt on US mint.

The mismatch between users and system design bias arose because of the change in a previously assumed target user group. The newly emerged user group can differ based on expertise and values. For example, incorporating automated checkout technology like kiosks in a system catering to new tech-savvy users where the majority demographic is actually comprised of senior citizens, could create bias against the senior citizens by belittling them based on their ability to adapt to new system changes.

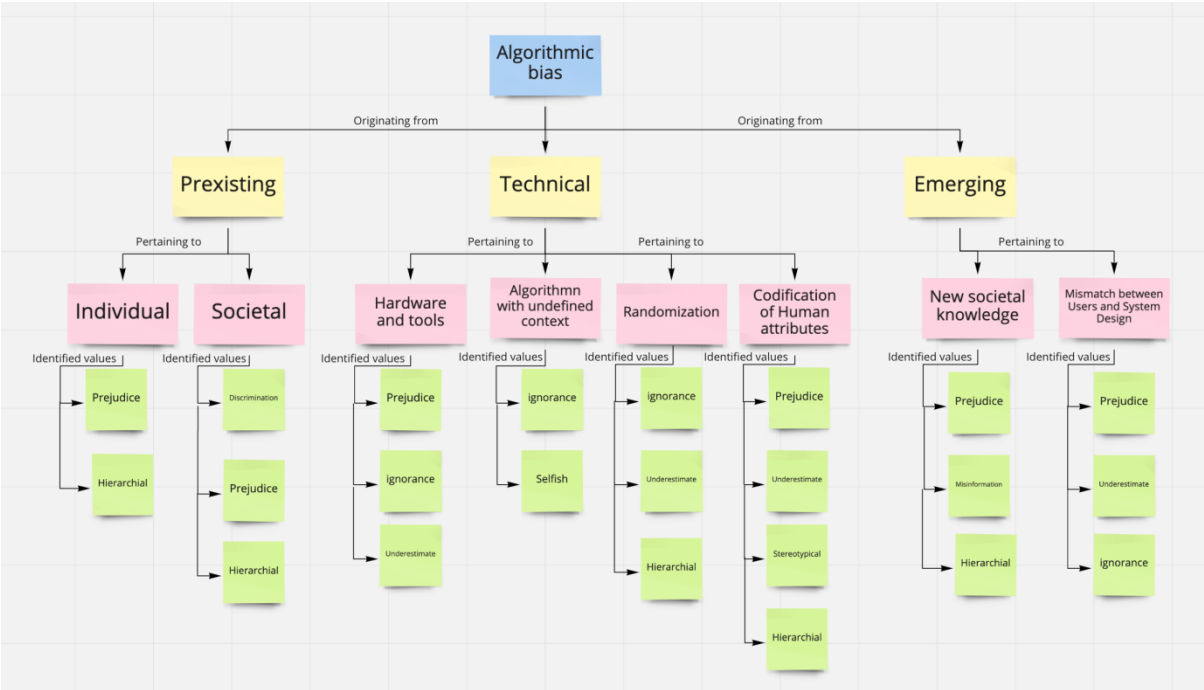

Figure 3: The map illustrates the distinction of algorithmic bias based on the origin and different contexts. Further, values are identified in each defined category.

Source: author

2.1.3 What values should be chosen?

The conceptualization and breakdown of bias in different contexts pertaining to technology facilitated the value mapping process. (Figure 3) Exploring all the identified values in the context of AFRT in law enforcement needs to be clustered in a thematic way which will facilitate the detailed analysis and mitigation of lacking value in the system. Based on the recurring identified value, prejudice is determined as a chosen value in the context of algorithm.

Algorithmic prejudice can be viewed as the relationships between protected traits and other criteria that are the source of algorithmic discrimination. For example, in a hypothetical situation, police use an AFRT to identify a convict based on the surveillance footage. Assuming the convict is a black male, here the protected trait will be black males. Other criteria might include time and location. Coincidentally, during the crime, the surveillance footage also captures black female, who is an innocent individual. AFRT's low accuracy in identifying faces based on gender and race could tag the innocent black female as a convict (figure 4 and figure 5). Now, even if the protected feature were to be modified from black male to black person, the other criteria might not be enough to accurately pinpoint the individual.

2.2 Empirical

To substantiate this claim, empirical research on AFRT's accuracy was done. The empirical model of inquiry necessitates a more situated understanding of the socio-technical system, allowing for not only the observation of stakeholders' usage and

appropriation patterns, but also the determination of whether the values envisioned during the design process are realized, amended, or subverted. Empirical research can also assist in answering questions such as, how is the value identified in the conceptual model being violated in AFRT and law enforcement?

The initial studies led by researchers at MIT (Buolamwini & Gebru, 2018) revealed the error rates for darker-skinned women were as high as 35 percent, 12 percent for darker-skinned males, 7% for lighter-skinned women, and less than 1% for lighter-skinned men.

Moreover, further research and audits conducted by the researchers of gender shades shed light on amazon's recognition technology accuracy rate being as low as 68.6 percent on the black women (figure 4). The dataset used for the research consisted of people looking straight into the camera, but in real life, surveillance footage can be more disoriented resulting in even more reduced accuracy.

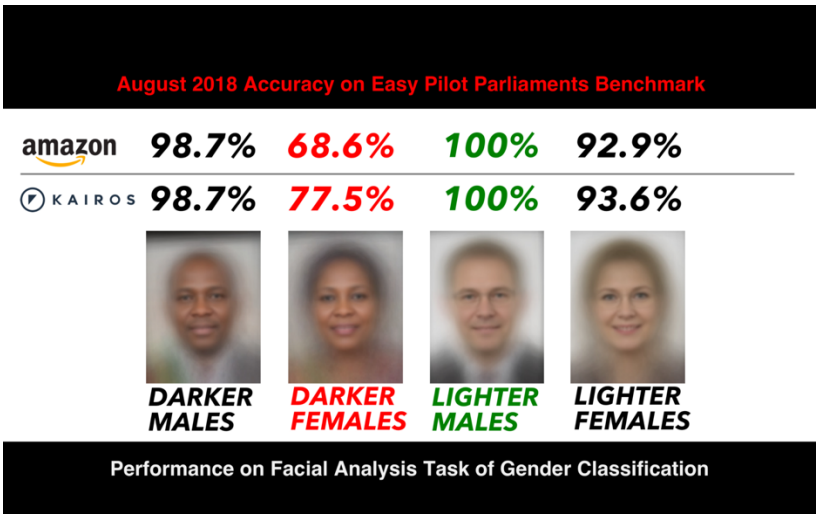

Figure 4: Comparison of ARFT accuracy in identifying faces based on gender and race.

Source: Buolamwini J., (2019). Response: Racial and Gender bias in Amazon Rekognition — Commercial AI System for Analyzing Faces. Retrieved from https://medium.com/@Joy.Buolamwini/response-racial-and-gender-bias-in-amazon-rekognition-commercial-ai-system-for-analyzing-faces

Figure 5: ARFT recognizing famous celebrity Oprah Winfrey as male with 76.7 percent accuracy claim.

Source: Buolamwini J., (2019). Response: Racial and Gender bias in Amazon Rekognition — Commercial AI System for Analyzing Faces. Retrieved from https://medium.com/@Joy.Buolamwini/response-racial-and-gender-bias-in-amazon-rekognition-commercial-ai-system-for-analyzing-faces

In their further research, the accuracy of AFRT was so maligned that the software identified black females as males which substantiate the claim demonstrated in algorithmic prejudice. (figure 4 and figure 5)

2.3 Technological investigation

Scholars have linked many of the hazards and probable negative effects of AFRT to prejudice as a derived value from algorithmic bias in law enforcement. More specifically, the insights gained through value conceptualization and empirical evidence show that we should evaluate values in context to acquire a thorough understanding of what is at stake and how we might transform our concerns into concrete design needs. It's especially vital to evaluate the AFRT misidentification implications in this scenario.

Value hierarchy table(VHT) (figure 6) can be used to visualize and aid designers in translating abstract values into technical design requirements in order to transform

the abstract value into technical design requirements., we present a concrete example of the tool below. The tool aids in comprehending the frequently abstract pathways for value translation. Values are translated into tangible design needs through these channels, and vice versa.

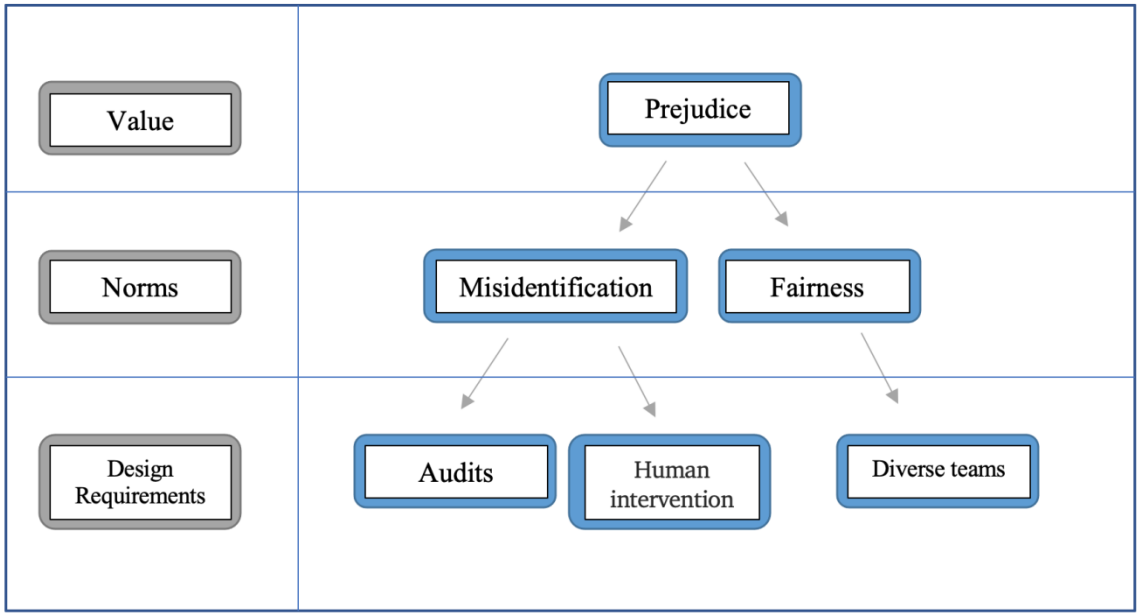

Figure 6: Value hierarchy table the systematic breakdown of contextual value and translation of the same using identified norms to deduce design requirements. *Source: Author*

The term "value" refers to the attractiveness of behaviour, whereas "norm" refers to a certain trait that allows for the accomplishment of that desired quality. For example, in figure 7, the value under scrutiny is prejudice, yet prejudice can be interpreted in a variety of ways by various people. As a result, standards such as misidentification and fairness emphasize intended value in discrete fragments. By misidentification, the research calls into doubt the accuracy of the AFRT, and by fairness, it calls into question the system's design construct. The derived design requirements can be further analysed to mitigate and system limitations and flaws. VHT can be used iteratively by designers, developers, and stakeholders towards the ethical design of the system.

2.3.1 Audit

Auditing can reveal a plethora of issues in the system. In the case of AFRT in law enforcement, it revealed the qualities of datasets used to test the algorithm. The study conducted by Gender shade audit reveals the disparities between organizations who considered auditing their existing AFRT algorithm versus organizations like Amazon who refused to audit their AFRT. (figure 7)

Table 1: Overall Error on Pilot Parliaments Benchmark, August 2018 (%)

Company	All	Females	Males	Darker	Lighter	DF	DM	LF	LM
Target Corporations									
Face ++	1.6	2.5	0.9	2.6	0.7	4.1	1.3	1.0	0.5
MSFT	0.48	0.90	0.15	0.89	0.15	1.52	0.33	0.34	0.00
IBM	4.41	9.36	0.43	8.16	1.17	16.97	0.63	2.37	0.26
Non-Target Corporations									
Amazon	8.66	18.73	0.57	15.11	3.08	31.37	1.26	7.12	0.00
Kairos	6.60	14.10	0.60	11.10	2.80	22.50	1.30	6.40	0.00

Table 2: Overall Error Difference Between August 2018 and May 2017 PPB Audit (%)

Company	All	Females	Males	Darker	Lighter	DF	DM	LF	LM
Face ++	-8.3	-18.7	0.2	-13.9	-3.9	-30.4	0.6	-8.5	-0.3
MSFT	-5.72	-9.70	-2.45	-12.01	-0.45	-19.28	-5.67	-1.06	0.00
IBM	-7.69	-10.74	-5.17	-14.24	-1.93	-17.73	-11.37	-4.43	-0.04

Figure 7: The table reveals the reduction inaccuracies of target corporations who employed auditing vs the corporation that did not implement auditing.

Source: Buolamwini & Gebru, (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of Machine Learning Research. Conference on Fairness, Accountability, and Transparency

2.3.2 Diverse team

AI reflects racism, sexism, and other prejudices based on the data supplied to it by its very creators. According to a 2016 study of over 20,000 companies in 91 countries, organizations with more female executives were more lucrative (Ellemers & Floor, 2016). It's crucial to have emotional intelligence on your team, whether you're a

woman, a man, or a nonbinary person. At the absolute least, the AI's development team should look like the people who will use it. In the case of AFRT, this could entail holding codesign workshops and contextual inquiries with designers from various backgrounds in order to have a deeper understanding of the system's consequences.

2.3.3 Human intervention in scrutinizing the result of the algorithm

The capacities of human specialists and automated algorithms were compared in case studies published by the National Institute of Standards and Technology (NIST) in January 2020. A facial identification test was conducted by forensic facial examiners and professional facial reviewers. A leading algorithm produced findings that were identical to those of the most skilled people. Nonetheless, the highest level of AFRT accuracy was only achieved through machine-human collaboration. (Charles, 2020)

3 Conclusion

This paper discusses how the VSD framework can be used to mitigate and provide design requirements by taking contextual value in the AFRT employed in law enforcement. The study also reveals the ability of VSD in revealing completely new values, understanding which values are at stake for a specific application, how they should be understood in that specific case, and how they might translate into design requirements. The limitations of VSD are visible when the issues arising from algorithmic bias in general call for consideration of multiple values in context to understand complete awareness of what is at stake and what might be done to translate our concerns into feasible design requirements.VSD as an approach helps to mitigate very specific values, but a broader systematic analysis of values might not be beneficial. Moreover, the accuracy of VSD could be unreliable as it depends on the quality and quantity of stakeholders inputs and empirical investigations.

4 Citations

[1] Buolamwini, J.& Gebru, T. (2018). *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*. Proceedings of Machine Learning Research. Conference on Fairness, Accountability, and Transparency

[2] Quach, K. (2018). Facial recognition software easily IDs white men, but error rates soar for black women. The Register. Retrieved 2019-07-21.

[3] Winner,L. (1980). Do Artifacts Have Politics? Daedalus, 109(1), 121 136.

[4] Simon, J. & Wong, P.-H. & Rieder, G. (2020). Algorithmic bias and the Value Sensitive Design approach. Internet Policy Review, 9(4)

[5] Ainsley F., (2018,July 30). Dark Pattern Design — It's Downright Unethical & Irresponsible. [Web log post]. Retrieved from https://uxplanet.org/dark-pattern-design-its-downright-unethical-irresponsible

[6] Barcoas,S.,& Selbst, A. D. (2016). Big data's disparate impact. California Law Review, 104(3), 671–732.

[7] Friedman, B., Kahn, P. H., & Borning, A. (2006). Value Sensitive Design and Information Systems.

[8] Friedman, B., & Hendry, D. G. (2019). Value Sensitive Design: Shaping Technology with Moral Imagination. MIT Press.

[9] Flanagan, M., Howe, D. C., & Nissenbaum, H. (2008). Embodying Values in Technology: Theory and Practice. In J. Hoven & J. Weckert (Eds.), Information Technology and Moral Philosophy (pp. 322–353). Cambridge University Press.

[10] Zhou, Marrian (2018, June 26). Orlando stops using Amazon's controversial facial recognition tech. CNET.

[11] Keane, Sean (2018,June 22). Amazon employees protest sale of face recognition software to police. CNET.

[12] Umbrello, S.(2022). Meaningful human control over smart home systems: a value sensitive design approach. Hum. Ment J. Philos. Stud. 13(37), 40–65 (2020)

[13] Freidman B., & Nissenbaum H. (1996). Bias in Computer Systems. ACM Transactions on Information Systems, Vol. 14, No. 3.

[14] Lauret J., (2019,August 16). Amazon’s sexist AI recruiting tool: how did it go so wrong. [Web blog post]. Retrieved from https://becominghuman.ai/amazons-sexist-ai-recruiting-tool-how-did-it-go-so-wrong

[15] Carlo S., (2018, July 6) We’ve got to stop the Met Police’s dangerously authoritarian facial recognition surveillance.[Web blog post] Retrieved from https://metro.co.uk/2018/07/06/weve-got-to-stop-the-met-polices-dangerously-authoritarian-facial-recognition-surveillance-7687833/

[16] Umbrello, S. & van de Poel, I., (2021).Mapping value sensitive design onto AI for social good principles. AI Ethics 1, 283–296

[17] Ellemers N., & Floor R.,(2016). Diversity in work groups, Current Opinion in Psychology,Volume 11, 2016, Pages 49-53,ISSN 2352-250X,

[18] Charles H. R., (2020), Facial Recognition Technology. (Part III): Ensuring Commercial Transparency & Accuracy.